

• 论 著 • DOI:10. 3969/j. issn. 1672-9455. 2026. 11. 012

基于随机森林模型预测乙型肝炎病毒表面抗原定量检测结果及临床应用*

李潇涵,唐卓芸,李冬冬[△]

四川大学华西医院实验医学科,四川成都 610041

摘要:**目的** 构建随机森林模型预测乙型肝炎病毒表面抗原(HBsAg)定量检测结果,优化检测流程,减少复测次数与试剂消耗。**方法** 基于 2025 年 1 月 1 日至 6 月 30 日于四川大学华西医院实验医学科进行 HBsAg 定量检测的 37 530 例样本,回顾性收集同一时间段同一患者的肝功能指标、HBV DNA、年龄、性别等特征,使用合成少数类过采样技术处理样本不平衡,通过网格搜索与阈值调优(侧重高于检出限样本召回率)训练模型。**结果** 该研究构建的随机森林模型对高于检出限样本召回率达 98.40%,总体准确率达 82.08%。该模型后单月可节约试剂 344 人份。**结论** 随机森林模型能有效预测 HBsAg 结果,提升检测效率,为类似滴度检测场景提供参考。

关键词:随机森林; 乙型肝炎病毒表面抗原; 预测模型; 合成少数类过采样技术; 临床应用
中图法分类号:R512.62 **文献标志码:**A **文章编号:**1672-9455(2026)11-1511-07

Prediction of hepatitis B surface antigen quantitative detection results based on random forest model and its clinical application*

LI Xiaohan, TANG Zhuoyun, LI Dongdong[△]

Department of Laboratory Medicine, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China

Abstract: Objective A random forest model was constructed to predict the quantitative detection results of hepatitis B virus surface antigen (HBsAg), optimize the detection process, and reduce the number of retests and reagent consumption. **Methods** Based on 37 530 samples of HBsAg quantitative detection in the Department of Experimental Medicine of West China Hospital of Sichuan University from January 1 to June 30, 2025, liver function indicators, HBV DNA, age, gender and other characteristics of the same patient in the same time period were retrospectively collected. The synthetic minority oversampling technique was used to deal with the sample imbalance, and the model was trained by grid search and threshold tuning (focusing on the recall rate of samples above the detection limit). **Results** The recall rate of the model constructed in the study for samples above the detection limit was 98.40%, and the overall accuracy rate was 82.08%. After application of the model, 344 reagents could be saved in a single month. **Conclusion** Random forest model can effectively predict HBsAg results, improve the detection efficiency, and provide reference for similar titer detection scenarios.

Key words: random forest; hepatitis B virus surface antigen; prediction model; synthetic minority oversampling technique; clinical application

乙型肝炎病毒(HBV)感染是我国肝硬化、肝细胞癌等慢性肝病的主要病因。据统计,我国慢性 HBV 感染者约 7 500 万例,2022 年一般人群 HBV 表面抗原(HBsAg)流行率为 5.6%,因人口基数大,相关疾病负担仍然沉重^[1-2]。HBsAg 定量检测对预测疾病进展、评估抗病毒疗效及预后有重要价值^[3-6]。然而,不同临床分期的感染者 HBsAg 水平差异显著,免疫耐受期可达 10⁶ IU/mL,而常用化学发光法的检测上

限通常为 100~250 IU/mL^[7-9],因此,在实际检测中常需对样本进行稀释。前期研究发现通过历史结果比较,1 年内同一患者 HBsAg 定性结果阴性或 HBsAg 定量结果低于检出限时不进行 HBsAg 稀释检测,可避免勾状效应导致错误结果,此方法可有效节约检测成本,缩短检测周转时间^[10]。但该优化方法仅依赖简单规则,效率有待提高。

* 基金项目:新发突发与重大传染病防控国家科技重大专项(2025ZD01906800)。
作者简介:李潇涵,男,主管技师,主要从事传染病筛查和实验室信息系统应用相关方向的研究。 [△] 通信作者,E-mail:jiangxili1219@163.com。
引用格式:李潇涵,唐卓芸,李冬冬. 基于随机森林模型预测乙型肝炎病毒表面抗原定量检测结果及临床应用[J]. 检验医学与临床,2026, 23(11):1511-1517.

随机森林作为一种集成学习方法,通过构建多棵决策树进行预测,具有良好的预测性能与抗过拟合能力,适用于多变量复杂数据的分析^[11-13]。尽管机器学习在预测 HBV 感染^[14-15]或清除^[16]上已有报道,但将此类模型用于指导具体的、单次的实验室检测操作的研究仍较匮乏。本研究基于罗氏 Cobas e801 检测平台,针对其检测流程中因钩状效应需对所有样本预先稀释、对超出范围者再进行原液检测的现状,结合前期研究基础及传统的模型构建步骤^[17-18],尝试建立随机森林模型预测 HBsAg 检测结果,以期优化检测流程,提升检测效率。

1 资料与方法

1.1 一般资料 选取 2025 年 1 月 1 日至 6 月 30 日于四川大学华西医院实验医学科进行 HBsAg 定量检测样本共 37 530 例(除外根据前期研究^[10]即结合既往史已经判断的样本)作为研究对象,其中 HBsAg 结果低于检出限(0.05 IU/mL)8 426 例,高于检出限 29 104 例。

1.2 方法 使用 Python(PyCharm)及 Scikit-learn 库进行随机森林建模、参数调优与性能评估。

1.2.1 数据收集 收集所有患者的基线资料,包括年龄(AGE)、性别(SEX)、就诊科室(DEPT)、临床诊断(DIAGNOSIS);收集与 HBsAg 采样时间间隔 ≤ 2 d 的天冬氨酸转氨酶(AST)、丙氨酸转氨酶(ALT)、总胆红素(TB)、直接胆红素(DB)、 γ -谷氨酰转氨酶(GGT)、碱性磷酸酶(ALP)、总胆汁酸(TBA)、清蛋白(ALB)、总蛋白(TP)等项目的检测结果, ≤ 7 d 的 HBV DNA(包括普通和高精度定量)检测结果。

1.2.2 数据处理 (1)特征编码:针对就诊科室,包括“感染”“传染”的科室赋值为 1,其余赋值为 0;针对临床诊断,包括“乙型肝炎”或“乙型病毒性肝炎”或“乙肝”的诊断赋值为 1,其余诊断赋值为 0;性别使用系统原值,男性赋值为 1,女性赋值为 2;HBV DNA 优先取高精定量检测结果,扩增阴性赋值为 0;所有数据中 $>X$ 均赋值为 X , $<X$ 赋值为 $X/10$ 。(2)缺失值与异常值处理:缺失值使用该列数据中位数进行填充,连续性变量使用 1%和 99%分位数截断处理。(3)数据标准化:对连续变量使用 Z-score 标准化处理。(4)数据划分:按 8:2 随机分为训练集与测试集。训练集中低于检出限样本 6 741 例,高于检出限 23 283 例;测试集中低于检出限 1 685 例,高于检出限 5 821 例。

1.2.3 样本类别不平衡处理 由于训练集中 HBV DNA 低于检出限与高于检出限的样数量差距较大,因此使用合成少数类过采样技术(SMOTE)对训练集中低于检出限的样本进行过采样,使两类样本量均为 23 283 例。

1.2.4 模型优化与调优 基于前期研究^[10]及医学相关随机森林模型的构建^[19-21],临床应用时需要尽量保证高于检出限样本的预测更加精准,减少该类样本的人工复测,需尽可能提高高于检出限样本的召回率,

因此结合下列参数在 SMOTE 过采样的训练集中进行分层交叉验证获得最优的超参数组合并计算交叉验证分数:(1)随机森林中树的数量(100,200,300);(2)树的最大深度(10,15,20,None);(3)内部节点再划分所需的最小样本数(2,5,10);(4)叶子节点最少样本数(1,2,4);(5)寻找最佳分割时考虑的特征数(sqr,log2,0.3,0.5);(6)使用 Bootstrap(True,False):如果为 True,则每棵树使用 Bootstrap 进行(有放回抽样)构建;如果为 False,则使用整个数据集构建每棵树。

在此基础上测试集对模型进行微调,在 0.1~0.9 范围内以 0.05 为步长调整分类阈值,以最大化高于检出限样本的 F1 分数。

1.2.5 模型性能评估 (1)使用基尼指数评估模型目标特征的重要性。(2)判别性能评估:绘制受试者工作特征(ROC)曲线并计算曲线下面积(AUC)及其 95%CI,评估模型区分“高于检出限”与“低于检出限”样本的整体能力;同时基于 ROC 曲线分析及侧重高召回率的临床需求,确定最终分类阈值。(3)预测性能评估:绘制校准曲线,通过比较模型预测的“高于检出限”概率与测试集中的实际阳性比例,计算概率预测均方误差(Brier Score)与对数损失(Log loss),结合模型预测概率分布相关指标(均值、中位数、标准差)评估模型预测概率的准确性。(4)在最终分类阈值下,应用最佳参数获得测试集的精确率(Precision)、召回率(Recall)、准确率(Accuracy)、F1 分数,评估最终分类效果。(5)使用原始训练集通过交叉验证计算交叉验证准确率,使用模型预测测试集结果,并计算测试集预测准确率共同评估模型泛化能力。(6)绘制学习曲线,结合过采样后的训练集的准确率与测试集准确率比较,共同评估该模型的过拟合程度。

1.2.6 模型外部验证与应用评估 收集 2025 年 8 月 1—31 日四川大学华西医院实验医学科进行 HBsAg 定量检测样本共 6 030 例作为外部验证集,预测时使用训练集的阈值,并且针对外部验证数据的缺失值,使用训练时的中位数进行填充,绘制 ROC 曲线,与测试集 AUC 结果进行比较,进一步评估模型泛化能力,同时计算预测准确率,结合预测结果评估可节约试剂人份数。

1.3 统计学处理 采用 SPSS27.0 统计软件进行数据处理与统计分析。符合正态分布的计量资料以 $\bar{x} \pm s$ 表示,2 组间比较采用独立样本 t 检验。不符合正态分布的计量资料以 $M(P_{25}, P_{75})$ 表示,2 组间比较采用 Mann-Whitney U 检验。计数资料以例数或百分率表示,组间比较采用 χ^2 检验。以 $P < 0.05$ 为差异有统计学意义。

2 结果

2.1 不同临床特征的 HBsAg 定量检测患者模型构建与外部验证数据集资料比较 HBsAg 结果低于检出限的人群 HBV DNA 数据缺失率、肝功指标缺失率

均高于 HBsAg 结果高于检出限的人群,差异均有统 计学意义($P<0.05$)。见表 1。

表 1 不同临床特征的 HBsAg 定量检测患者模型构建与外部验证数据集资料比较[$\bar{x}\pm s$ 或 $M(P_{25},P_{75})$ 或 $n(\%)$]

临床特征	模型构建		外部验证		$t/\chi^2/H$	P_1	$t/\chi^2/H$	P_2
	高于检出限 ($n=29\ 104$)	低于检出限 ($n=8\ 426$)	高于检出限 ($n=4\ 334$)	低于检出限 ($n=1\ 696$)				
年龄(岁)	45.22±11.34	43.94±15.20	44.95±11.19	42.17±12.39	1.437	0.151	4.219	<0.001
就诊科室 ^a	8 746(30.1)	7 205(85.5)	1 442(33.3)	1 448(85.4)	18.473	<0.001	0.020	0.795
诊断 ^a	8 423(28.9)	7 878(93.5)	1 329(30.7)	1 586(93.5)	5.424	<0.001	0.001	0.960
男性	17 422(59.9)	4 079(48.4)	2 561(59.1)	790(46.6)	0.931	0.137	1.893	0.169
ALT(U/L)	24(20,32)	22(18,29)	23(19,28)	24(19,34)	0.040	0.841	2.007	0.157
AST(U/L)	24(16,37)	21(14,36)	22(16,34)	24(15,41)	4.160	0.248	0.755	0.385
TB(μmol/L)	12.8(9.9,16.8)	12.4(9.3,16.8)	13.3(10.1,17.4)	14.1(10.7,19.8)	1.168	0.280	6.349	0.070
DB(μmol/L)	3.3(2.3,4.2)	3.1(2.3,4.2)	3.4(2.7,4.8)	3.8(2.9,5.4)	71.741	<0.001	31.219	<0.001
GGT(U/L)	21.0(14.0,39.0)	23.0(14.8,47.0)	22.0(14.0,34.0)	27.0(16.0,58.0)	7.098	0.046	0.244	0.621
ALP(U/L)	81.0(66.0,98.0)	82.0(65.8,102.0)	82.0(56.0,102.0)	86.0(72.0,104.0)	6.827	0.054	11.633	0.004
TBA(μmol/L)	3.1(1.8,5.7)	2.5(1.3,4.9)	3.1(1.8,5.6)	3.1(1.8,5.9)	17.759	<0.001	7.147	0.045
HBV DNA(IU/mL)	10.0(10.0,10.0)	10.0(10.0,10.0)	10.0(10.0,10.0)	10.0(10.0,10.0)	3.634	0.057	2.843	0.092
ALB(g/L)	46.10±3.61	46.74±3.02	46.56±3.63	47.05±3.13	-6.020	<0.001	-3.939	<0.001
TP(g/L)	73.48±4.94	74.43±4.32	73.81±5.20	74.25±4.38	2.392	<0.001	-2.041	<0.001
数据缺失								
肝功能指标 ^b	1 457(5.0)	2 447(29.0)	290(6.7)	565(33.3)	12.330	<0.001	21.632	<0.001
HBV DNA	3 364(11.6)	7 055(83.7)	840(19.38)	1 356(80.0)	14.332	<0.001	210.033	<0.001
其余特征	0(0.00)	0(0.00)	0(0.00)	0(0.00)	—	—	—	—

注：^a表示统计特征编码为 1 的科室及诊断；^b表示包括 ALT、AST、TB、DB、GGT、ALP、TBA； P_1 为模型构建数据集与外部验证数据集中，高于检出限组各特征比较的 P 值； P_2 为模型构建数据集与外部验证数据集中，低于检出限组各特征比较的 P 值；—表示无数据。

2.2 最优超参数与阈值 最优超参数组合为：树数量 300、最大深度无限制、叶节点最小样本数 1、内部节点最小样本数 5、特征数 0.5、Bootstrap 为 False。最佳分类阈值为 0.1，即模型投票≥10%即判定为“高于检出限”。此设定下，“高于检出限”样本的召回率为 0.984 0，F1 分数为 0.894 9。

2.3 模型性能评估

2.3.1 基于基尼指数的模型特征重要性 重要性前 5 的特征依次为：临床诊断(Gini=34.22)、就诊科室(Gini=13.69)、HBV DNA 结果(Gini=10.60)、年龄(Gini=5.28)、GGT(Gini=4.95)。见图 1。

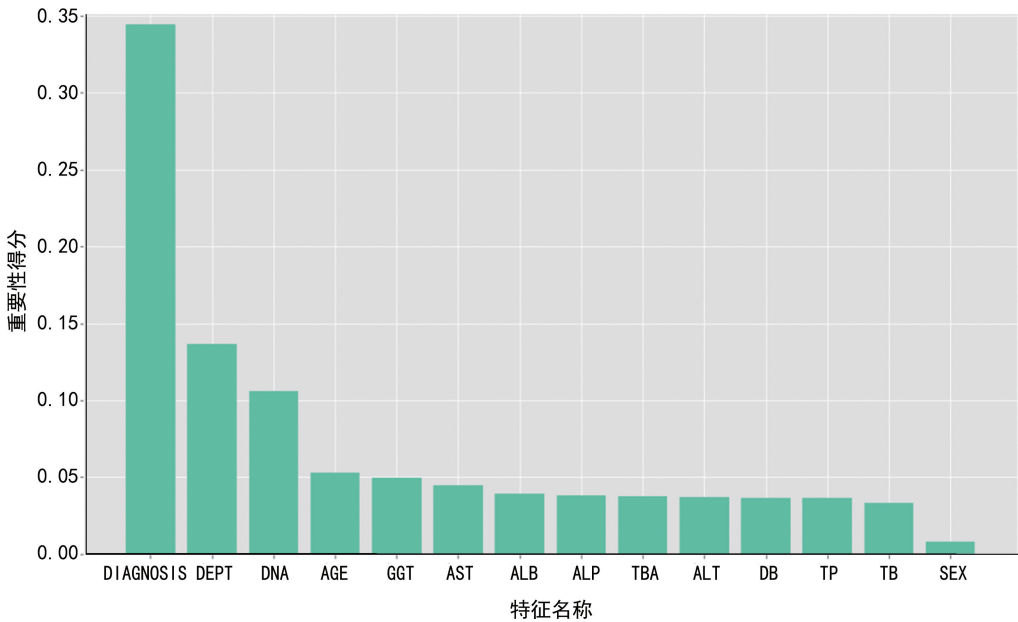


图 1 基于基尼指数的模型特征重要性

2.3.2 判别性能与 ROC 曲线分析 测试集中,模型区分“高于检出限”与“低于检出限”样本的 AUC 为 0.919 4(95%CI:0.912 1~0.926 6)。见图 2。

2.3.3 预测校准曲线及预测概率分布 校准曲线的 Brier 分数为 0.094 5,Log Loss 为 0.379 8,通过曲线拟合模型校准曲线的斜率为 0.863 8,截距为 0.137 5, R^2 为 0.999 4。预测概率分布结果显示均值为 0.740,中位数为 0.948,标准差为 0.334。见图 3。

2.3.4 基于阈值的分类性能 当使用以上超参数和阈值时该模型的测试集总体准确率为 82.08%,高于检出限样本的精确率为 82.06%,召回率为 98.40%,F1 分数为 0.894 9;低于检出限样本的精确率为 82.32%,召回率为 25.70%,F1 分数为 0.391 7。见表 2。使用 SMOTE 过采样后交叉验证分数为

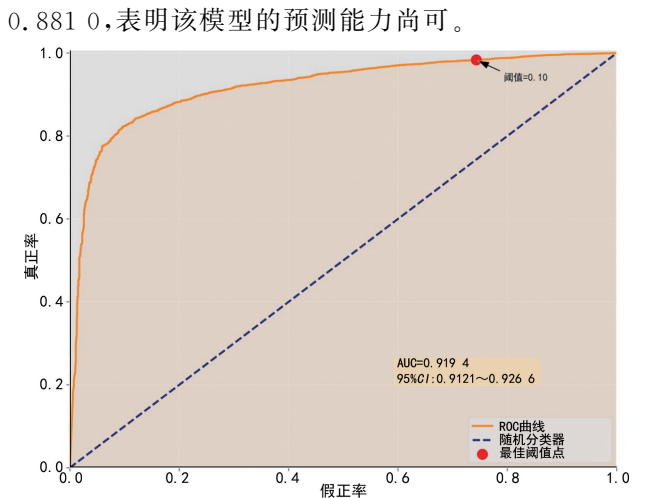
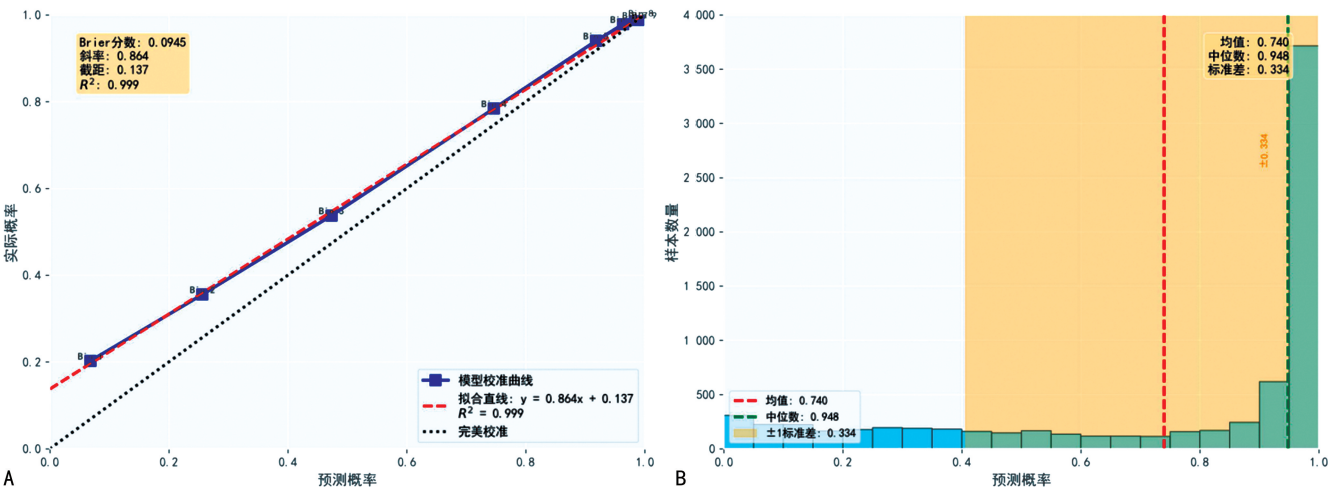


图2 测试集 ROC 曲线及最佳阈值点标记



注:A 为校准曲线;B 为预测概率分布。

图3 模型校准曲线及预测概率分布

表 2 最优超参数下测试集性能

分类	样本量	预测正确	预测错误	精确率(%)	召回率(%)	F1 分数	准确率(%)
低于检出限样本	1 685	433	1 252	82.32	25.70	0.391 7	—
高于检出限样本	5 821	5 728	93	82.06	98.40	0.894 9	—
总样本	7 506	6 161	1 345	82.08	82.08	0.802 8	82.08

注:—表示无数据。

2.3.5 泛化能力 使用原始训练集计算获得的交叉验证准确率为 $0.869\ 9\pm0.000\ 7$,测试集预测准确率为 0.863 2,2 个准确率接近,表明该模型的泛化能力较好。

2.3.6 过拟合程度 过采样后数据的训练准确率为 $0.984\ 4\pm0.000\ 5$,与测试集预测准确率 0.863 2 相比,存在一定差异,学习曲线结果也表明存在一定程度的过拟合,但整体得分较高,过拟合尚可接受。见图 4。

2.4 模型外部验证及应用评估 使用该模型(阈值为 0.1)进行外部验证,其中低于检出限样本 1 696 例,高于检出限样本 4 334 例。ROC 曲线结果显示外

部验证的 AUC 为 0.910 9(95%CI:0.902 8~0.919 2),与测试集 AUC(0.919 4)接近,表明模型泛化能力较好。见图 5。

使用训练时的阈值(0.1)进行预测时,低于检出限样本的预测准确率为 25.18%(427/1 696);高于检出限样本的预测准确率为 98.08%(4 251/4 334),总体预测准确率为 77.58%(4 678/6 030)。基于 HBV DNA 在低检出限人群的高缺失率,外部验证时若去除 HBV DNA 检测结果,低于检出限人群预测正确率为 25.59%(434/1 696),高于检出限人群预测正确率为 97.23%(4 214/4 334)。

按常规检测流程,高于检出限样本以只进行一次

稀释后检测计算,低于检出限的样本需先进行稀释检测,再进行不稀释检测,总试剂消耗理论值为 7 726。按以上预测结果,高于检出限的样本误判 83 例,每例需人工进行一次复测,低于检出限样本正确预测 427 例,每例可减少一次复测,该模型理论单月可节约试剂 344 人份,约占总试剂消耗的 4.45%(344/7 726)。

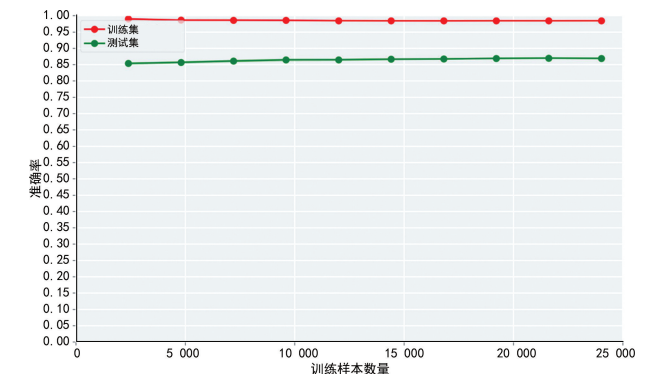


图 4 模型学习曲线

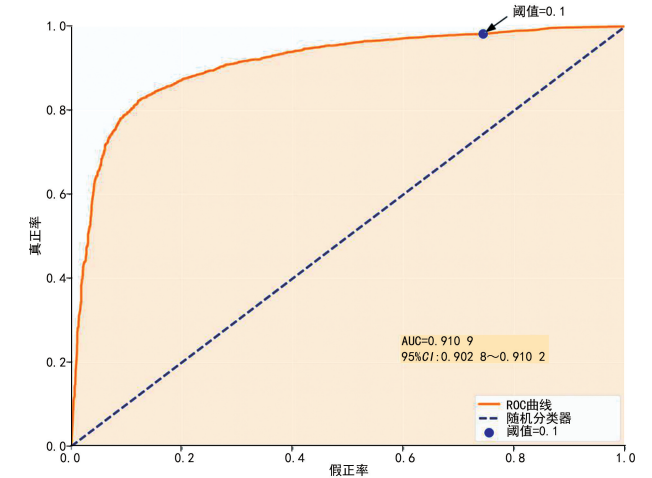


图 5 外部验证集 ROC 曲线及应用阈值点标记

3 讨 论

本研究的核心创新点在于将机器学习模型:随机森林,作为一种前瞻性工具应用于 HBsAg 定量检测的流程优化,目的是节约试剂成本,提升检测效率。这一思路为临床检验实验室的工作流程改进提供了新的视角。

本研究人群的数据特征与特征重要性分析结果高度吻合,高于检出限人群主要就診科室为感染科和传染科,而低于检出限就診人群主要就診科室是便民门诊;同时高于检出限人群的诊断与“乙型肝炎”高度相关,低于检出限人群的诊断则并不明确,主要为“便民自助”。这一差异深刻反映了 HBsAg 水平与 HBV 感染临床管理重心之间的内在联系:活动性感染或需要监测的患者更倾向于在专科就診并拥有明确诊断,其 HBsAg 水平往往较高;而常规筛查或非活动性携带者则可能通过便民门诊检测,其 HBsAg 水平常处于低值或阴性范围,模型有效地捕捉并学习了这一临床实践模式。

外部验证 AUC 为 0.910 9(95%CI:0.902 8~0.919 2),与测试集 AUC(0.919 4)极为接近;高于检出限样本的召回率在外部验证中为 98.08%,与测试集的 98.40%相差无几。尽管总体准确率在外部验证中略有下降(77.57% 对比 82.08%),但对高效识别需稀释样本(高召回率)的影响极小,这种稳定性说明模型在不同时间段的样本上具有可靠的判别能力。外部验证结果与测试集表现高度一致,表明模型具有良好的泛化能力。

外部验证数据集中部分实验室指标(如 DB、ALP、TBA 等)在 2 组间分布与训练集存在统计学差异,这可能源于人群随时间推移的自然波动或检测批次差异。然而模型仍保持优异性能,进一步体现了随机森林对数据微小变化的鲁棒性。

在 HBsAg 定量检测的实际工作中,由于钩状效应的存在,实验室通常需要对所有样本进行预稀释,并对超出检测范围的样本进行复测^[7-9]。这一传统流程确保了检测质量,但不可避免地导致了试剂与时间的双重消耗。本研究通过模型预测,实现了对检测流程的“精准预判”。本研究模型对测试集中高于检出限样本的召回率高达 98.40%,表明绝大多数真正需要稀释检测的样本能被准确识别。尽管该模型对低于检出限样本的识别能力(召回率为 25.70%)有限,但鉴于本研究的首要目标是避免对高浓度样本的漏判,大量样本仍可通过检测平台既有的复检规则按照原来的检测流程自动完成复测,从而减少手动操作,有效控制检测周转时间。外部验证结果显示,单月可节约试剂 344 人份,占总消耗的 4.45%,充分验证了该策略在提升实验室运营效率方面的实际价值。

本研究是前期研究的深化与拓展。此前尝试通过逻辑规则预测稀释模式,取得了 5 个月理论节约 1 476 人份试剂的效果^[10]。相比之下,随机森林模型能够捕捉变量间复杂的非线性关系,提供了更强大的预测能力。与国内外研究多将机器学习聚焦于智能审核^[22]或辅助疾病诊断^[23-26]的模型不同,本研究将机器学习应用于“检验”这一核心环节,直接优化检测操作本身,提升临床检测效率。

特征重要性分析结果也与临床认知具有高度一致性。临床诊断和就診科室成为最重要的预测因子,这与 HBsAg 水平同 HBV 感染自然史密切相关的观点相符^[27]。HBV DNA 作为病毒复制活跃度的直接标志物,其重要性亦得到证实^[3-4,27-28]。部分肝功能指标(如 GGT)的重要性也与既往研究中 HBsAg 与肝功能关联性的报告相互证实^[29]。这些发现共同表明,模型有效学习了 HBsAg 水平与患者临床状态的内在生物学联系。

对于 HBV DNA 等特征的缺失值,研究采用了基于训练集中位数填充的预处理方法。这是一种基于工程实践考虑的常见处理方式,其有效性依赖于“缺

失并非完全随机,且分布与训练集相似”的假设。然而,结合本研究的临床场景,该方法具有其合理性:从临床逻辑看,HBsAg 低于检出限的人群通常不属于活动性 HBV 感染重点关注人群,临床医生未常规开展 HBV DNA 检测导致其呈现高比例的“结构性缺失”。使用训练集中位数(10 IU/mL)即赋值后的常见检测下限填充,实质上是将此类样本归类为“典型的阴性或低病毒载量模式”,符合临床实践认知。外部验证表明,无论是否包括 HBV DNA 特征,模型对核心目标(高于检出限样本)的召回率均维持在极高水平(98.08% 与 97.23%)。这说明模型决策高度依赖“临床诊断”与“就诊科室”等更强特征,HBV DNA 的辅助作用相对有限,因此简单填充策略对主要目标影响可控。多重插补技术等更复杂的方法可能提供更优估计,但会显著增加系统复杂性^[30]。本研究以解决临床实际需求与简便部署为首要原则,当前策略在性能与复杂度间取得了良好平衡。

模型展现了良好的泛化能力。测试集与外部验证集的 AUC 非常接近,且交叉验证准确率与测试集准确率相差无几。学习曲线虽提示存在一定过拟合,但测试性能仍然稳健,说明模型在未见数据上具有良好的预测能力。这得益于随机森林算法本身的抗过拟合特性以及研究中采用的特征工程、SMOTE 处理和交叉验证调参等措施。

该模型所代表的“预测-优化”策略,在检验医学领域具有广阔的推广前景。许多类似的定量或半定量检测,如梅毒 TRUST 滴度检测、隐球菌抗原滴度检测等,其检测流程同样严重依赖操作者对滴度范围的预判。这些场景均可借鉴本研究思路,通过构建适当的预测模型,在保证质量的前提下,显著缩短检测周转时间,实现资源的精准配置。

当然,本研究存在若干局限性。为确保模型的临床可解释性与易实施性,所选特征主要基于文献报道和临床经验^[27-29],未能纳入所有潜在影响因素,如 HBeAg 状态^[31]、药物治疗史等,这可能在一定程度上限制了模型的性能上限。研究数据来源于单一中心,尽管样本量大,但模型的普适性仍需在不同人群和检测平台上进行外部验证。模型预测依赖于肝功能指标和 HBV DNA 的先期获取,这对实验室信息系统的数据整合能力提出了要求。

综上所述,本研究成功构建了一个基于随机森林的 HBsAg 定量检测结果预测模型,并将其创新性地应用于检测流程的优化,且该模型能在确保检测质量的同时,有效提升检测效率、降低试剂消耗。更重要的是,本研究为临床检验实验室应对各类需要经验预判的检测场景如滴度测定,提供了一种较为客观的、基于数据驱动的流程优化新范式。

利益冲突 所有作者均声明不存在利益冲突。
作者贡献 李潇涵:提出研究思路,设计研究方

案,数据收集、清洗、预处理及特征工程,完成随机森林模型的构建、调优、评估与验证,进行统计分析 & 结果解读,撰写论文初稿,完成图表的绘制与论文修改;唐卓芸:参与部分文献调研与资料整理,对论文内容与表述进行修改;李冬冬:指导研究设计与数据分析思路,对论文结构、内容与表述进行修改与润色,审定最终版本。

参考文献

[1] 中华预防医学会感染性疾病防控分会,中华预防医学会促进消除病毒性肝炎工作委员会,林炳亮,等. 成人乙型肝炎病毒感染筛查、检测及管理专家建议[J]. 临床肝胆病杂志, 2024, 40(8): 1543-1550.

[2] COOKE G S, FLOWER B, CUNNINGHAM E, et al. Progress towards elimination of viral hepatitis: a lancet gastroenterology & hepatology commission update[J]. Lancet Gastroenterol Hepatol, 2024, 9(4): 346-365.

[3] 中华医学会, 中华医学会杂志社, 中华医学会全科医学分会, 等. 慢性乙型肝炎基层诊疗指南(实践版·2020)[J]. 中华全科医师杂志, 2021, 20(3): 281-289.

[4] 中华医学会感染病学分会, 中华医学会肝病学分会. 慢性乙型肝炎防治指南(2019 年版)[J/CD]. 中华实验和临床感染病杂志(电子版), 2019, 13(6): 441-466.

[5] LAZAREVIC I, BANKO A, MILJANOVIC D, et al. Hepatitis B surface antigen isoforms: their clinical implications, utilisation in diagnosis, prevention and new antiviral strategies [J]. Pathogens, 2024, 13(1): 46.

[6] LEOWATTANA W, LEOWATTANA P, LEOWATTANA T. Quantitative hepatitis B core antibody and quantitative hepatitis B surface antigen: novel viral biomarkers for chronic hepatitis B management [J]. World J Hepatol, 2024, 16(4): 550-565.

[7] PARK J, BAE T, CHO Y, et al. Analytical performance of the sysmex HISCL HBsAg assay and comparison with the roche elecsys HBsAg II quant assay in the quantification of hepatitis B surface antigen[J]. Medicina (Kaunas), 2021, 57(12): 1307.

[8] YAN Y, SUN H Z, CHANG L, et al. Circulating immune complexes and mutations of HBsAg are associated with the undetectable HBsAg in anti-HBs and HBeAg positive occult hepatitis B virus infection[J]. Front Microbiol,

- 2022,13;1063616.
- [9] LEE J,LEE S Y,CHO Y G,et al. Accuracy validation of the elecsys HBsAg II quant assay and its utility in resolving equivocal qualitative HBsAg results[J]. Medicina (Kaunas), 2023, 59(3):443.
- [10] 李潇涵,李冬冬. 乙肝表面抗原定量检测稀释模式预测及应用[J]. 临床检验杂志, 2023,41(5): 385-387.
- [11] 高梦媛,李梦,董丽. 基于机器学习算法构建呼吸机相关性肺炎预测模型的研究进展[J]. 全科护理, 2025,23(15):2855-2859.
- [12] SUO P C,WANG X Y,KOU W L,et al. WeNet-RF: an automatic classification model for financial reimbursement budget items[J]. PLoS One, 2025,20(4):e0321056.
- [13] BECKER T,ROUSSEAU A J,GEUBBELMANS M,et al. Decision trees and random forests [J]. Am J Orthod Dentofacial Orthop, 2023, 164(6):894-897.
- [14] WANG Y,DU Z,LAWRENCE W R,et al. Predicting hepatitis B virus infection based on health examination data of community population[J]. Int J Environ Res Public Health, 2019, 16(23):4842.
- [15] YAO N,LIU Y,XU J W,et al. Identification of associated risk factors for serological distribution of hepatitis B virus via machine learning models[J]. BMC Infect Dis, 2024,24(1):66.
- [16] YAN F,TANG F,CHEN J,et al. Exploring using HBsAg to predict interferon treatment course to achieve clinical cure in chronic hepatitis B patients:a clinical study[J]. Front Immunol, 2024,15:1528758.
- [17] BISWAS A,SARAN I,WILSON F P. Introduction to supervised machine learning [J]. Kidney360, 2021,2(5):878-880.
- [18] 南锐,刘燕丽,万祁阳,等. 基于机器学习的糖尿病预测和因素分析[J]. 智能计算机与应用, 2025,15(6):140-145.
- [19] 邹新亮,郑万香,何国祥,等. 基于随机森林算法预测肾病综合征患者的心血管风险[J]. 重庆医学, 2022,51(3):393-397.
- [20] WANG H,SONG Y,BI H. Optimizing public health management with predictive analytics: leveraging the power of random forest [J]. Front Big Data, 2025,8:1574683.
- [21] 贾辰,高岩,解夕黎,等. 基于 5 种机器学习算法构建 ICU 患者耐碳青霉烯类鲍曼不动杆菌感染风险预测模型[J]. 中华医院感染学杂志, 2025, 35(17):2586-2591.
- [22] 刘骏龙,许晓泓,王华珍,等. 基于机器学习的医学检验智能审核模型[J]. 中国数字医学, 2019, 14(3):53-55.
- [23] 王艳,贺鑫,梁玉芳,等. 基于常规检验指标预测血流感染的机器学习模型构建与评价[J]. 中华医院感染学杂志, 2025,35(10):1542-1548.
- [24] 黄莺,周颖,宋云霄,等. 基于机器学习算法构建血常规指标肺结核诊断模型[J]. 检验医学, 2024,39(7):668-672.
- [25] 王小科,晏峻峰. 基于随机森林的帕金森疾病诊断模型构建研究[J]. 计算机技术与发展, 2023, 33(4):154-160.
- [26] 杨琴,刘秀燕,史玉媛. 随机森林模型用于分析系统性红斑狼疮患者的预后因素[J]. 临床医药实践, 2025,34(9):665-669.
- [27] 朱霞,唐红. HBsAg 定量与 HBV DNA 载量在 HBV 感染自然史不同阶段的临床意义[J]. 四川大学学报(医学版), 2019,50(5):771-774.
- [28] 邢同京,赵坤宇,李文涛,等. 慢性 HBV 感染者病毒 DNA 水平与患者肝组织炎症损伤程度的相关性研究[J]. 中华肝脏病杂志, 2023,31(9): 954-960.
- [29] GÜNAL Ö,BARUT ŞETIKAN İ,et al. Relation between serum quantitative HBsAg, ALT and HBV DNA levels in HBeAg negative chronic HBV infection[J]. Turk J Gastroenterol, 2014,25 Suppl 1:142-146.
- [30] ARCHER L,KOSHIARIS C,LAY-FLURRIE S,et al. Development and external validation of a risk prediction model for falls in patients with an indication for antihypertensive treatment: retrospective cohort study[J]. BMJ, 2022,379: e070918.
- [31] 张晓晶,华乐,金巧菲,等. 慢性 HBV 感染者 HBV RNA 与乙肝病毒学标志物及肝功能指标的相关性[J]. 中华医院感染学杂志, 2024, 34(11):1616-1620.

(收稿日期:2025-10-16 修回日期:2026-02-22)

(编辑:陈晶 李菲菲)